

The Minimum Regularized Covariance Determinant estimator

Kris Boudt*

Solvay Business School, Vrije Universiteit Brussel
Faculty of Economics and Business, Vrije Universiteit Amsterdam

Peter J. Rousseeuw

Department of Mathematics, KU Leuven

Steven Vanduffel

Solvay Business School, Vrije Universiteit Brussel

Tim Verdonck

Department of Mathematics, KU Leuven

January 26, 2017

Abstract

The Minimum Covariance Determinant (MCD) approach estimates the location and scatter matrix using the subset of given size with lowest sample covariance determinant. Its main drawback is that it cannot be applied when the dimension exceeds the subset size. We propose the Minimum Regularized Covariance Determinant (MRCD) approach, which differs from the MCD in that the subset-based covariance matrix is a convex combination of a target matrix and the sample covariance matrix. A data-driven procedure sets the weight of the target matrix, so that the regularization is only used when needed. The MRCD estimator is defined in any dimension, is well-conditioned by construction and preserves the good robustness properties of the MCD. We prove that so-called concentration steps can be performed to reduce the MRCD objective function, and we exploit this fact to construct a fast algorithm. We verify the accuracy and robustness of the MRCD estimator in a simulation study and illustrate its practical use for outlier detection and regression analysis on real-life high-dimensional data sets in chemistry and criminology.

Keywords: Breakdown point; High-dimensional data; Regularization; Robust covariance estimation.

*This research has benefited from the financial support of the Flemish Science Foundation (FWO) and project C16/15/068 of Internal Funds KU Leuven. We are grateful to Yukai Yang for his initial assistance to this work. We also thank Dries Cornilly, Christophe Croux, Sebastiaan Höppner, Stefan Van Aelst and Marjan Wauters for their constructive comments.

1 Introduction

The Minimum Covariance Determinant (MCD) method (Rousseeuw, 1984, 1985) is a highly robust estimator of multivariate location and scatter. Given an $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ with $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, its objective is to find h observations whose sample covariance matrix has the lowest possible determinant. Here $h < n$ is fixed. The MCD estimate of location is then the average of these h points, whereas the scatter estimate is a multiple of their covariance matrix. Consistency and asymptotic normality of the MCD estimator have been shown by Butler et al. (1993) and Cator and Lopuhaä (2012). The MCD has a bounded influence function (Croux and Haesbroeck, 1999) and has the highest possible breakdown value (i.e. 50%) when $h = \lfloor (n + p + 1)/2 \rfloor$ (Lopuhaä and Rousseeuw, 1991). The MCD approach has been applied to various fields such as chemistry, finance, image analysis, medicine, and quality control, see e.g. the review paper of Hubert et al. (2008).

A major restriction of the MCD approach is that the dimension p must satisfy $p < h$ for the covariance matrix of any h -subset to be non-singular. In fact, for accuracy of the estimator it is often recommended to take $n > 5p$, e.g. in Rousseeuw et al. (2012). This limitation creates a gap in the availability of high breakdown methods for so-called “fat data”, in which the number of rows (observations) is small compared to the number of columns (variables). To fill this gap we propose a modification of the MCD to make it applicable to high dimensions. The basic idea is to replace the subset-based covariance by a regularized covariance estimate, defined as a weighted average of the sample covariance of the h -subset and a predetermined positive definite target matrix. The proposed Minimum Regularized Covariance Determinant (MRCD) estimator is then the regularized covariance based on the h -subset which makes the overall determinant the smallest.

In addition to its availability for high dimensions, the main features of the MRCD estimator are that it preserves the good breakdown properties of the MCD estimator and is well-conditioned by construction. Since the estimated covariance matrix is guaranteed to be invertible it is suitable for computing robust distances, and for linear discriminant analysis and graphical modeling (Öllerer and Croux, 2015). Furthermore, we will generalize the C-step theorem of Rousseeuw and Van Driessen (1999) by showing that the objective function is reduced when concentrating the h -subset to the h observations with the smallest robust distance computed from the regularized covariance. This C-step theorem forms the theoretical basis for the proposed fast MRCD estimation algorithm.

The remainder of the paper is organized as follows. In Section 2 we introduce the MRCD covariance estimator and discuss its properties. Section 3 proposes a practical and fast algorithm for the MRCD. The extensive simulation study in Section 4 confirms the good properties of the method. Section 5 uses

the MRCD estimator for outlier detection and regression analysis on real data sets from chemistry and criminology. The main findings and suggestions for further research are summarized in the conclusion.

2 From MCD to MRCD

Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ be a dataset in which $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ denotes the i -th observation ($i = 1, \dots, n$). The observations are stored in the $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$. We assume that most of them come from an elliptical distribution with location $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$. The remaining observations can be arbitrary outliers, and we do not know beforehand which ones they are. The problem is to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ despite the outliers.

2.1 The MCD estimator

The MCD approach searches for an h -subset of the data (where $n/2 \leq h < n$) whose sample covariance matrix has the lowest possible determinant. Clearly, the subset size h affects the efficiency of the estimator as well as its robustness to outliers. For robustness, $n - h$ should be at least the number of outliers. When many outliers could occur one may set $h = \lceil 0.5n \rceil$. Typically one sets $h = \lceil 0.75n \rceil$ to get a better efficiency. Throughout the paper, H denotes a set of h indices reflecting the observations included in the subset, and \mathcal{H}_h is the collection of all such sets. For a given H in \mathcal{H}_h we denote the corresponding $h \times p$ submatrix of \mathbf{X} by \mathbf{X}_H . Throughout the paper, we use the term h -subset to denote both H and \mathbf{X}_H interchangeably. The mean and sample covariance matrix of \mathbf{X}_H are then

$$\mathbf{m}_\mathbf{X}(H) = h^{-1} \mathbf{X}_H' \mathbf{1}_h \quad (1)$$

$$\mathbf{S}_\mathbf{X}(H) = h^{-1} (\mathbf{X}_H - \mathbf{m}_\mathbf{X}(H))' (\mathbf{X}_H - \mathbf{m}_\mathbf{X}(H)) \quad (2)$$

The MCD approach then aims to minimize the determinant of $\mathbf{S}_\mathbf{X}(H)$ among all $H \in \mathcal{H}_h$:

$$H_{MCD} = \operatorname{argmin}_{H \in \mathcal{H}_h} (\det(\mathbf{S}_\mathbf{X}(H))^{1/p}) \quad (3)$$

where we take the p -th root of the determinant for numerical reasons. Note that the p -th root of the determinant of the covariance matrix is the geometric mean of its eigenvalues; SenGupta (1987) calls it the standardized generalized variance. The MCD can also be seen as a multivariate least trimmed squares estimator in which the trimmed observations have the largest Mahalanobis distance with respect to the sample mean and covariance of the h -subset (Agulló et al., 2008).

The MCD estimate of location \mathbf{m}_{MCD} is defined as the average of the h -subset, whereas the MCD scatter estimate is given as a multiple of its sample covariance matrix:

$$\mathbf{m}_{MCD} = \mathbf{m}_{\mathbf{X}}(H_{MCD}) \quad (4)$$

$$\mathbf{S}_{MCD} = c_{\alpha} \mathbf{S}_{\mathbf{X}}(H_{MCD}) \quad (5)$$

where c_{α} is a consistency factor such as the one given by Croux and Haesbroeck (1999), and depends on the trimming percentage $\alpha = (n - h)/n$. Butler et al. (1993) and Cator and Lopuhaä (2012) prove consistency and asymptotic normality of the MCD estimator, and Lopuhaä and Rousseeuw (1991) show that it has the highest possible breakdown value (i.e., 50%) when $h = \lfloor (n + p + 1)/2 \rfloor$. Accurately estimating a covariance matrix requires a sufficiently high number of observations. A rule of thumb is to require $n > 5p$ (Rousseeuw and Van Zomeren, 1990; Rousseeuw et al., 2012). When $p > h$ the MCD is ill-defined since all $\mathbf{S}_{\mathbf{X}}(H)$ have zero determinant.

2.2 The MRCD estimator

We will generalize the MCD estimator to high dimensions. As is common in the literature, we first standardize the p variables. For this we compute the median of each variable and stack them in a location vector $\nu_{\mathbf{X}}$. We also estimate the scale of each variable by the Qn estimator of Rousseeuw and Croux (1993), and put these scales in a diagonal matrix $\mathbf{D}_{\mathbf{X}}$. The standardized observations are then

$$\mathbf{u}_i = \mathbf{D}_{\mathbf{X}}^{-1}(\mathbf{x}_i - \nu_{\mathbf{X}}) \quad (6)$$

This disentangles the location-scale and correlation problems, as in Boudt et al. (2012).

In a second step, we use a predetermined and well-conditioned symmetric and positive definite target matrix \mathbf{T} . Following Won et al. (2013), we call such a matrix well-conditioned if the condition number (i.e., the ratio between the largest and smallest eigenvalues) is at most 1000. We also use a scalar weight coefficient ρ , henceforth called the regularization parameter. We then define the regularized covariance matrix of an h -subset H of the standardized data \mathbf{U} as

$$\mathbf{K}(H) = \rho \mathbf{T} + (1 - \rho)c_{\alpha} \mathbf{S}_{\mathbf{U}}(H) \quad (7)$$

where $\mathbf{S}_{\mathbf{U}}(H)$ is as defined in (2) but for \mathbf{U} , and c_{α} is the same consistency factor as in (5).

It will be convenient to use the singular value decomposition $\mathbf{T} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$ where $\mathbf{\Lambda}$ is the diagonal matrix holding the eigenvalues of \mathbf{T} and \mathbf{Q} is the orthogonal matrix holding the corresponding eigenvectors. We can then rewrite the regularized covariance matrix $\mathbf{K}(H)$ as

$$\mathbf{K}(H) = \mathbf{Q}\mathbf{\Lambda}^{1/2}[\rho \mathbf{I} + (1 - \rho)c_{\alpha} \mathbf{S}_{\mathbf{W}}(H)]\mathbf{\Lambda}^{1/2}\mathbf{Q}' \quad (8)$$

where the $n \times p$ matrix \mathbf{W} consists of the transformed standardized observations $\mathbf{w}_i = \mathbf{\Lambda}^{-1/2} \mathbf{Q}' \mathbf{u}_i$. It follows that $\mathbf{S}_{\mathbf{W}}(H) = \mathbf{\Lambda}^{-1/2} \mathbf{Q}' \mathbf{S}_U(H) \mathbf{Q} \mathbf{\Lambda}^{-1/2}$.

The MRCD subset H_{MRCD} is defined by minimizing the determinant of the regularized covariance matrix $\mathbf{K}(H)$ in (8):

$$H_{MRCD} = \underset{H \in \mathcal{H}_h}{\operatorname{argmin}} \left(\det(\mathbf{K}(H))^{1/p} \right) . \quad (9)$$

Since \mathbf{T} , \mathbf{Q} and $\mathbf{\Lambda}$ are fixed, H_{MRCD} can also be written as

$$H_{MRCD} = \underset{H \in \mathcal{H}_h}{\operatorname{argmin}} \left(\det(\rho \mathbf{I} + (1 - \rho) c_\alpha \mathbf{S}_{\mathbf{W}}(H))^{1/p} \right) . \quad (10)$$

Once H_{MRCD} is determined, the MRCD location and scatter estimates of the original data matrix \mathbf{X} are defined as

$$\mathbf{m}_{MRCD} = \nu_{\mathbf{X}} + \mathbf{D}_{\mathbf{X}} \mathbf{m}_U(H_{MRCD}) \quad (11)$$

$$\mathbf{K}_{MRCD} = \mathbf{D}_{\mathbf{X}} \mathbf{Q} \mathbf{\Lambda}^{1/2} [\rho \mathbf{I} + (1 - \rho) \mathbf{S}^*(H_{MRCD})] \mathbf{\Lambda}^{1/2} \mathbf{Q}' \mathbf{D}_{\mathbf{X}} \quad (12)$$

in which

$$\mathbf{S}^*(H_{MRCD}) = \mathbf{D}_{\mathbf{W}}^{-1} \mathbf{S}_{\mathbf{W}}(H_{MRCD}) \mathbf{D}_{\mathbf{W}}^{-1} \quad (13)$$

and where $\mathbf{D}_{\mathbf{W}}$ in (13) is the diagonal matrix holding the square roots of the diagonal elements of $\mathbf{S}_{\mathbf{W}}(H_{MRCD})$. We use $\mathbf{S}^*(H_{MRCD})$ to rescale the diagonal elements of the final MRCD covariance matrix (12). When the target matrix \mathbf{T} is the identity, these diagonal elements become those of the initial scale estimates in $\mathbf{D}_{\mathbf{X}}$. Also note that the consistency factor c_α cancels in (12).

Because of the initial standardization step, the MRCD scatter estimate is location invariant and scale equivariant. This means that for any diagonal $p \times p$ matrix \mathbf{A} and any $p \times 1$ vector \mathbf{b} the MRCD scatter estimate $S(\mathbf{A}\mathbf{X} + \mathbf{b})$ equals $\mathbf{A}S(\mathbf{X})\mathbf{A}'$.

2.3 The MRCD precision matrix

The precision matrix is the inverse of the scatter matrix, and is needed for the calculation of robust MRCD-based Mahalanobis distances, for linear discriminant analysis, for graphical modeling (Öllerer and Croux, 2015), and for many other computations. The MRCD scatter matrix (12) is computationally convenient to invert by its construction, yielding the expression

$$\mathbf{K}_{MRCD}^{-1} = \mathbf{D}_{\mathbf{X}}^{-1} \mathbf{\Lambda}^{-1/2} \mathbf{Q}' [\rho \mathbf{I}_p + (1 - \rho) \mathbf{S}^*(H_{MRCD})]^{-1} \mathbf{Q} \mathbf{\Lambda}^{-1/2} \mathbf{D}_{\mathbf{X}}^{-1} . \quad (14)$$

By the Sherman-Morrison-Woodbury identity (Sherman and Morrison, 1950; Woodbury, 1950; Bartlett, 1951) this can also be written as

$$\mathbf{K}_{MRCD}^{-1} = \mathbf{D}_{\mathbf{X}}^{-1} \mathbf{\Lambda}^{-1/2} \mathbf{Q}' \left[\frac{1}{\rho} \mathbf{I}_p - \frac{1}{\rho^2} \frac{1-\rho}{h} \mathbf{Z} \left(\mathbf{I}_h + \frac{1-\rho}{h} \frac{1}{\rho} \mathbf{Z}' \mathbf{Z} \right)^{-1} \mathbf{Z}' \right]^{-1} \mathbf{Q} \mathbf{\Lambda}^{-1/2} \mathbf{D}_{\mathbf{X}}^{-1} \quad (15)$$

where \mathbf{Z} is the $n \times p$ matrix holding the standardized observations

$$\mathbf{z}_i = \mathbf{D}_{\mathbf{U}}^{-1/2} (\mathbf{u}_i - \mathbf{m}_{\mathbf{U}}(H_{MRCD})) I(i \text{ belongs to } H_{MRCD}) \quad (16)$$

where $I(\cdot)$ is the indicator function, hence $\mathbf{S}^*(H_{MRCD}) = h^{-1} \mathbf{Z} \mathbf{Z}'$.

Note that the MRCD should not be confused with the Regularized Minimum Covariance Determinant (RMCD) estimator of Croux et al. (2012). The latter assumes sparsity of the precision matrix, and maximizes the penalized log-likelihood function of each h -subset by the GLASSO algorithm of Friedman et al. (2008). The repeated application of GLASSO is time-consuming.

2.4 Choice of target matrix and calibration of ρ

The MRCD estimate depends on two quantities: the target matrix \mathbf{T} and the regularization parameter ρ . For the target matrix \mathbf{T} on \mathbf{U} we can take the identity matrix; relative to the original data \mathbf{X} this is the diagonal matrix with the robustly estimated univariate scales on the diagonal. Depending on the application, we can also take a non-diagonal target matrix \mathbf{T} . When this matrix is estimated in a first step, it should be robust to outliers in the data. A reasonable choice is to compute a rank correlation matrix of \mathbf{U} , which incorporates some of the relation between the variables. When we have reasons to suspect an equicorrelation structure, we can set \mathbf{T} equal to

$$\mathbf{R}_c = c \mathbf{J}_p + (1 - c) \mathbf{I}_p \quad (17)$$

with \mathbf{J}_p the $p \times p$ matrix of ones, \mathbf{I}_p the identity matrix, and $-1/(p-1) < c < 1$ to ensure positive definiteness. The parameter c in the equicorrelation matrix (17) can be estimated by averaging robust correlation estimates over all pairs of variables, under the constraint that the determinant of \mathbf{R}_c is above a minimum threshold value.

When the regularization parameter ρ equals zero $\mathbf{K}(H)$ becomes the sample covariance $\mathbf{S}_{\mathbf{U}}(H)$, and when ρ equals one $\mathbf{K}(H)$ becomes the target. We require $0 \leq \rho \leq 1$ to ensure that $\mathbf{K}(H)$ is positive definite (as it is a convex combination of positive definite matrices), hence invertible.

In practice, we recommend a data-driven approach which sets ρ at the lowest nonnegative value for which $\rho \mathbf{I} + (1 - \rho) c_{\alpha} \mathbf{S}_{\mathbf{W}}(H)$ is well-conditioned. This is easy to implement, as we only need to compute

the eigenvalues λ of $c_\alpha \mathbf{S}_W(H)$ once, since the eigenvalues of $\rho \mathbf{I} + (1 - \rho)c_\alpha \mathbf{S}_W(H)$ equal

$$\rho + (1 - \rho)\lambda. \quad (18)$$

Note that by this heuristic we only use regularization when needed. Indeed, if $\mathbf{S}_W(H)$ is well-conditioned, the heuristic sets ρ equal to zero. Also note that the eigenvalues in (18) are at least ρ , so the smallest eigenvalue of the MRCD scatter estimate is bounded away from zero. Therefore the MRCD scatter estimator has a 100% implosion breakdown value, compared to the 50% implosion breakdown value of the MCD.

3 An algorithm for the MRCD estimator

A naive algorithm for the optimization problem (9) would be to compute $\det(\mathbf{K}(H))$ for every possible h -subset H . However, for realistic sample sizes this type of brute force evaluation is infeasible.

The original MCD estimator (3) has the same issue. The current solution for the MCD consists of either selecting a large number of randomly chosen initial subsets (Rousseeuw and Van Driessen, 1999) or starting from a smaller number of deterministic subsets (Hubert et al., 2012). In either case one iteratively applies so-called C-steps. The C-step of MCD improves an h -subset H_1 by computing its mean and covariance matrix, and then puts the h observations with smallest Mahalanobis distance in a new subset H_2 . The C-step theorem of Rousseeuw and Van Driessen (1999) proves that the covariance determinant of H_2 is lower than or equal to that of H_1 , so C-steps lower the MCD objective function.

We will now generalize this theorem to regularized covariance matrices.

Theorem 1. *Starting from a h -subset H_1 , one can compute $\mathbf{m}_1 = \frac{1}{h} \sum_{i \in H_1} \mathbf{x}_i$ and $\mathbf{S}_1 = \frac{1}{h} \sum_{i \in H_1} (\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)'$. The matrix*

$$\mathbf{K}_1 = \rho \mathbf{T} + (1 - \rho) \mathbf{S}_1$$

is positive definite hence invertible, so we can compute

$$d_1(i) = (\mathbf{x}_i - \mathbf{m}_1)' \mathbf{K}_1^{-1} (\mathbf{x}_i - \mathbf{m}_1)$$

for $i = 1, \dots, n$. Let H_2 be an h -subset for which

$$\sum_{i \in H_2} d_1(i) \leq \sum_{i \in H_1} d_1(i) \quad (19)$$

and compute $\mathbf{m}_2 = \frac{1}{h} \sum_{i \in H_2} \mathbf{x}_i$, $\mathbf{S}_2 = \frac{1}{h} \sum_{i \in H_2} (\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)'$ and $\mathbf{K}_2 = \rho \mathbf{T} + (1 - \rho) \mathbf{S}_2$. Then

$$\det(\mathbf{K}_2) \leq \det(\mathbf{K}_1) \quad (20)$$

with equality if and only if $\mathbf{m}_2 = \mathbf{m}_1$ and $\mathbf{K}_2 = \mathbf{K}_1$.

Note that for $\rho = 0$ our Theorem 1 specializes to Theorem 1 of Rousseeuw and Van Driessen (1999), but with the weaker condition (19), thereby strengthening the result. The proof of Theorem 1 is given in Appendix A.

Making use of the generalized C-step we can now construct the actual algorithm. In the pseudocode we see that Step 3.1 determines the initial scatter estimates \mathbf{S}^i of \mathbf{W} in the same way as in the DetMCD algorithm of Hubert et al. (2012). In case an initial \mathbf{S}^i is (nearly) singular, step 3.2 regularizes it with parameter $\rho_\bullet = 0.1$ (experiments have indicated that in most cases the same results are obtained for $\rho_\bullet = 0.05$ and $\rho_\bullet = 0.15$). Based on these subsets we then set ρ in a conservative way to ensure the MRCO covariance computed on each of them is well-conditioned. The value of ρ for which the condition number of $\rho \mathbf{I} + (1 - \rho)c_\alpha \mathbf{S}_W(H)$ equals 1000 is obtained using line search and formula (18).

For each initial subset we then apply C-steps until the subset no longer changes, which typically requires only a few steps. Finally, out of the six resulting subsets we select the one with the lowest objective, and use it to compute our final location and scatter estimates according to (12).

MRCD algorithm

1. Compute the standardized observations \mathbf{u}_i as defined in (6) using the median and the Qn estimator for univariate location and scale.
 2. Perform the singular value decomposition of \mathbf{T} into $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$ where $\mathbf{\Lambda}$ is the diagonal matrix holding the eigenvalues of \mathbf{T} and \mathbf{Q} is the orthogonal matrix whose columns are the corresponding eigenvectors. Compute $\mathbf{w}_i = \mathbf{\Lambda}^{-1/2}\mathbf{Q}'\mathbf{u}_i$.
 3. Find the MRCD subset:
 - 3.1. Follow Subsection 3.2 in Hubert et al. (2012) to obtain initial location estimates \mathbf{m}^i and scatter estimates \mathbf{S}^i for $i = 1, \dots, 6$.
 - 3.2. If \mathbf{S}^i is not invertible, replace \mathbf{S}^i by $\rho_{\bullet} \mathbf{I} + (1 - \rho_{\bullet})\mathbf{S}^i$ where $\rho_{\bullet} = 0.1$.
 - 3.3. Determine the subsets H_0^i of \mathbf{W} containing the h observations with lowest Mahalanobis distance in terms of \mathbf{m}^i and \mathbf{S}^i .
 - 3.4. For each subset H_0^i , determine the smallest value of $0 \leq \rho_i < 1$ for which $\rho \mathbf{I} + (1 - \rho)c_{\alpha}\mathbf{S}_{\mathbf{W}}(H_0^i)$ is well-conditioned. Denote this value as ρ_i .
 - 3.5. Set $\rho = \max_i \rho_i$.
 - 3.6. For each initial subset H_0^i ($i = 1, \dots, 6$), repeat the generalized C-steps from Theorem 1 until convergence. Denote the resulting subsets as H^i .
 - 3.7. Let H_{MRCD} be the subset for which $\rho \mathbf{I} + (1 - \rho)c_{\alpha}\mathbf{S}_{\mathbf{W}}(H^i)$ has the lowest determinant among the six candidates.
 4. From H_{MRCD} compute the final MRCD location and scatter estimates as in (12).
-

4 Simulation study

We now investigate the empirical performance of the MRCD. In this section and the next we will use an equicorrelated target matrix as in (17). We estimate the equicorrelation parameter c as the average of all pairwise correlation estimates obtained using Kendall's tau, the robustness of which was studied in (Croux and Dehon, 2010). To guarantee positive definiteness, we impose that c is at least 0.1 above

$-1/(p-1)$. We compare the MRCD estimator to the OGK estimator of Maronna and Zamar (2002), which can also robustly estimate location and scatter in high dimensions but by itself does not guarantee the scatter matrix is well-conditioned. The OGK estimator is described in Appendix B.

Data generation setup. In the simulation experiment we generated $M = 500$ contaminated samples of size n from a p -variate normal distribution, with $n \times p$ taken as either 800×100 , 400×200 , 200×400 and 100×800 . In the first setting we generated data with mean zero and all variances equal to one (this is without loss of generality, given the equivariance properties of the estimators being considered). The correlation matrices were generated randomly following Agostinelli et al. (2015), henceforth ALYZ, to ensure that the performance is not tied to a particular choice of correlation matrix. The randomization was restricted to correlation matrices with condition number equal to 100.

Our second setup is the three-factor model used by Fan et al. (2008). Under this model \mathbf{x}_i is generated as $\mathbf{x}_i = \mathbf{b} \mathbf{f}_i + \boldsymbol{\varepsilon}_i$ where the 3-dimensional random vectors \mathbf{f}_i are normal random variables and the $p \times 3$ factor loadings \mathbf{b} are drawn from a multivariate normal distribution. The mean and covariance of each of the component distributions match those in Fan et al. (2008, Table 1). The p -dimensional independent error terms are also normal random variables, with zero means and standard deviations generated from a gamma distribution that is bounded from below by a threshold value.

To contaminate the data sets, we follow Maronna and Zamar (2002) and randomly replace $\lfloor \varepsilon n \rfloor$ observations by outliers along the eigenvector direction of Σ with smallest eigenvalue, since this is the direction where the contamination is hardest to detect. The distance between the outliers and the mean of the good data is denoted by k , which is set to 50 for medium-sized outlier contamination and to 100 for far outliers. We let the fraction of contamination ε be either 0% (clean data), 10% or 20%.

Evaluation setup. On each generated data set we run the MRCD with different subset sizes h , taken as 50%, 75%, 90% and 100% of the sample size n , and compare the results obtained when using $\rho = 0$ (MCD estimator) versus using the data-driven choice of ρ with the condition number at most 1000. We measure the inaccuracy of our scatter estimates \mathbf{S}_m compared to the true covariance Σ_m by their mean squared error (MSE) given by

$$MSE = \frac{1}{M} \frac{1}{p^2} \sum_{m=1}^M \sum_{k=1}^p \sum_{l=1}^p (\mathbf{S}_m - \Sigma_m)_{k,l}^2 .$$

Note that the true Σ_m differs across values of m when generating data according to ALYZ.

The results are reported in Tables 1 and 2, where the left panel shows the MSE and the right panel lists the average value of the data-driven ρ . The top panel shows the results in the absence of outlier

Table 1: Mean squared error of the MRCD and OGK scatter matrices, together with the average value of ρ , across 500 replications of the ALYZ data generating process.

	<i>MSE</i>				<i>Average value of ρ</i>			
	800x100	400x200	200x400	100x800	800x100	400x200	200x400	100x800
<i>Panel A: Clean data</i>								
$h = \lceil 0.5n \rceil$	0.0028	0.0052	0.0099	0.0194	0.0000	0.0048	0.0098	0.0254
$h = \lceil 0.5n \rceil, \rho = 0$	0.0028							
$h = \lceil 0.75n \rceil$	0.0019	0.0035	0.0066	0.0130	0.0000	0.0000	0.0077	0.0188
$h = \lceil 0.75n \rceil, \rho = 0$	0.0019	0.0035						
$h = \lceil 0.9n \rceil$	0.0015	0.0029	0.0055	0.0109	0.0000	0.0000	0.0069	0.0165
$h = \lceil 0.9n \rceil, \rho = 0$	0.0015	0.0029						
$h = n$	0.0013	0.0025	0.0050	0.0098	0.0000	0.0000	0.0065	0.0153
$h = n, \rho = 0$	0.0013	0.0025						
OGK	0.0014	0.0027	0.0056	0.0111				
<i>Panel B: 10% contamination at $k = 50$</i>								
$h = \lceil 0.5n \rceil$	0.0074	0.0110	0.0196	0.0381	0.0000	0.0036	0.0073	0.0187
$h = \lceil 0.5n \rceil, \rho = 0$	0.0074							
$h = \lceil 0.75n \rceil$	0.0056	0.0076	0.0132	0.0256	0.0000	0.0000	0.0057	0.0139
$h = \lceil 0.75n \rceil, \rho = 0$	0.0056	0.0076						
$h = \lceil 0.9n \rceil$	0.0048	0.0064	0.0111	0.0214	0.0000	0.0000	0.0051	0.0121
$h = \lceil 0.9n \rceil, \rho = 0$	0.0048	0.0064						
$h = n$	0.4523	0.2892	0.1507	0.0628	0.2186	0.3721	0.5460	0.7075
$h = n, \rho = 0$	0.7345	0.7262						
OGK	0.0058	0.0068	0.0102	0.0195				
<i>Panel C: 20% contamination at $k = 50$</i>								
$h = \lceil 0.5n \rceil$	0.0202	0.0231	0.0371	0.0700	0.0000	0.0029	0.0057	0.0144
$h = \lceil 0.5n \rceil, \rho = 0$	0.0202							
$h = \lceil 0.75n \rceil$	0.0167	0.0169	0.0254	0.0471	0.0000	0.0001	0.0045	0.0107
$h = \lceil 0.75n \rceil, \rho = 0$	0.0167	0.0169						
$h = \lceil 0.9n \rceil$	1.2634	0.7624	0.4352	0.2021	0.1763	0.3095	0.4738	0.6425
$h = \lceil 0.9n \rceil, \rho = 0$	1.8222	1.5617						
$h = n$	1.0937	0.6688	0.3305	0.1311	0.2574	0.4165	0.5897	0.7422
$h = n, \rho = 0$	1.9636	1.9417						
OGK	0.0234	0.0187	0.0206	0.0214				
<i>Panel D: 10% contamination at $k = 100$</i>								
$h = \lceil 0.5n \rceil$	0.0085	0.0122	0.0216	0.0419	0.0000	0.0034	0.0069	0.0178
$h = \lceil 0.5n \rceil, \rho = 0$	0.0085							
$h = \lceil 0.75n \rceil$	0.0065	0.0085	0.0146	0.0281	0.0000	0.0000	0.0054	0.0132
$h = \lceil 0.75n \rceil, \rho = 0$	0.0065	0.0085						
$h = \lceil 0.9n \rceil$	0.0057	0.0072	0.0122	0.0235	0.0000	0.0000	0.0049	0.0116
$h = \lceil 0.9n \rceil, \rho = 0$	0.0057	0.0072						
$h = n$	0.2767	0.1147	0.0397	0.0121	0.5427	0.7049	0.8273	0.9059
$h = n, \rho = 0$	1.2816	1.2672						
OGK	0.0066	0.0076	0.0126	0.0277				
<i>Panel E: 20% contamination at $k = 100$</i>								
$h = \lceil 0.5n \rceil$	0.0265	0.0290	0.0455	0.0858	0.0000	0.0026	0.0051	0.0130
$h = \lceil 0.5n \rceil, \rho = 0$	0.0265							
$h = \lceil 0.75n \rceil$	0.0224	0.0215	0.0314	0.0577	0.0000	0.0000	0.0041	0.0096
$h = \lceil 0.75n \rceil, \rho = 0$	0.0224	0.0215						
$h = \lceil 0.9n \rceil$	0.8509	0.3761	0.1394	0.0453	0.4740	0.6423	0.7812	0.8768
$h = \lceil 0.9n \rceil, \rho = 0$	3.0287	2.8113						
$h = n$	0.5555	0.2232	0.0751	0.0230	0.5879	0.7404	0.8508	0.9195
$h = n, \rho = 0$	3.1467	3.1499						
OGK	0.0333	0.0257	0.0273	0.0513				

Table 2: Mean squared error of the MRCD and OGK scatter matrices, together with the average value of ρ , across 500 replications of the linear factor model data generating process.

	MSE				Average value of ρ			
	800x100	400x200	200x400	100x800	800x100	400x200	200x400	100x800
<i>Panel A: Clean data</i>								
$h = [0.5n]$	0.0061	0.0110	0.0228	0.0431	0.0000	0.0126	0.0187	0.0389
$h = [0.5n], \rho = 0$	0.0061							
$h = [0.75n]$	0.0050	0.0086	0.0169	0.0320	0.0000	0.0003	0.0196	0.0384
$h = [0.75n], \rho = 0$	0.0050	0.0086						
$h = [0.9n]$	0.0046	0.0077	0.0143	0.0273	0.0000	0.0000	0.0214	0.0417
$h = [0.9n], \rho = 0$	0.0046	0.0077						
$h = n$	0.0043	0.0071	0.0124	0.0237	0.0000	0.0000	0.0250	0.0493
$h = n, \rho = 0$	0.0043	0.0071						
OGK	0.0037	0.0069	0.0169	0.0318				
<i>Panel B: 10% contamination at $k = 50$</i>								
$h = [0.5n]$	0.0078	0.0117	0.0219	0.0415	0.0000	0.0112	0.0188	0.0395
$h = [0.5n], \rho = 0$	0.0078							
$h = [0.75n]$	0.0065	0.0092	0.0163	0.0308	0.0000	0.0007	0.0196	0.0392
$h = [0.75n], \rho = 0$	0.0065	0.0092						
$h = [0.9n]$	0.0059	0.0082	0.0137	0.0253	0.0000	0.0000	0.0229	0.0471
$h = [0.9n], \rho = 0$	0.0059	0.0082						
$h = n$	0.0558	0.0512	0.0596	0.0691	0.6691	0.7899	0.8795	0.9370
$h = n, \rho = 0$	0.1676	0.1361						
OGK	0.0121	0.0180	0.0345	0.0436				
<i>Panel C: 20% contamination at $k = 50$</i>								
$h = [0.5n]$	0.0108	0.0131	0.0219	0.0413	0.0000	0.0111	0.0194	0.0399
$h = [0.5n], \rho = 0$	0.0108							
$h = [0.75n]$	0.0093	0.0105	0.0161	0.0295	0.0000	0.0032	0.0210	0.0425
$h = [0.75n], \rho = 0$	0.0093	0.0105						
$h = [0.9n]$	0.0917	0.0606	0.0648	0.0696	0.5961	0.7323	0.8417	0.9154
$h = [0.9n], \rho = 0$	0.3021	0.2089						
$h = n$	0.0753	0.0583	0.0629	0.0699	0.6985	0.8117	0.8940	0.9452
$h = n, \rho = 0$	0.3212	0.2510						
OGK	0.1642	0.1790	0.1566	0.0920				
<i>Panel D: 10% contamination at $k = 100$</i>								
$h = [0.5n]$	0.0112	0.0133	0.0230	0.0421	0.0000	0.0104	0.0175	0.0384
$h = [0.5n], \rho = 0$	0.0112							
$h = [0.75n]$	0.0098	0.0108	0.0171	0.0312	0.0000	0.0005	0.0184	0.0379
$h = [0.75n], \rho = 0$	0.0098	0.0108						
$h = [0.9n]$	0.0092	0.0098	0.0146	0.0259	0.0000	0.0000	0.0214	0.0447
$h = [0.9n], \rho = 0$	0.0092	0.0098						
$h = n$	0.0511	0.0529	0.0590	0.0688	0.8879	0.9358	0.9659	0.9830
$h = n, \rho = 0$	0.2994	0.2476						
OGK	0.0135	0.0166	0.0339	0.0476				
<i>Panel E: 20% contamination at $k = 100$</i>								
$h = [0.5n]$	0.0216	0.0178	0.0244	0.0431	0.0000	0.0101	0.0180	0.0377
$h = [0.5n], \rho = 0$	0.0216							
$h = [0.75n]$	0.0200	0.0151	0.0183	0.0309	0.0000	0.0027	0.0194	0.0401
$h = [0.75n], \rho = 0$	0.0200	0.0151						
$h = [0.9n]$	0.0705	0.0592	0.0614	0.0691	0.8495	0.9121	0.9529	0.9765
$h = [0.9n], \rho = 0$	0.5471	0.3844						
$h = n$	0.0662	0.0588	0.0612	0.0693	0.8985	0.9426	0.9699	0.9850
$h = n, \rho = 0$	0.5753	0.4408						
OGK	0.1348	0.1746	0.1480	0.1399				

contamination, i.e. $\varepsilon = 0\%$. In the lower panels we see the effects of 10% and 20% contamination in the data, for different values of k .

Discussion of results. Let us first study the relationship between the MRCD and the MCD. The latter is found in the rows labeled $\rho = 0$, and only exists in the settings where $p > h$. The MRCD instead uses the smallest value of $0 \leq \rho < 1$ for which the scatter matrix is well-conditioned, so when the MCD is well-conditioned the MRCD also obtains $\rho = 0$ and thus coincides with the MCD in that case. We see this in the tables when $n \times p$ equals 800×100 and $\varepsilon < (n - h)/n$. But when ε is larger, the MCD part becomes singular and hence more weight ρ is put on the target matrix, so the regularization makes the MRCD scatter matrix well-conditioned.

Interestingly, for $\varepsilon = 20\%$ of outliers the choice $h = \lceil 0.9n \rceil$ is too high, and in that case we see that the MRCD actually has a lower MSE for $k = 100$ than for $k = 50$. This is due to the fact that outliers further away induce higher ρ values.

The main alternative for the MRCD in high-dimensional scatter matrix estimation is the OGK method of Maronna and Zamar (2002). As can be seen in Appendix B, the OGK estimator does not result from optimizing an explicit objective function like the M(R)CD approach. Nevertheless it often works well in practice. The OGK estimator is listed in the bottom line of each panel in Tables 1 and 2. We see that the MSE of the MRCD with $h = \lceil 0.75n \rceil$ is roughly similar to that of the OGK for ALYZ data, whereas for the factor data the MRCD estimator substantially outperforms the OGK.

In conclusion, the simulation study confirms that the MRCD is a good method for estimating location and scatter in high dimensions. It only regularizes when needed. When h is less than p and the number of clean observations, the resulting ρ is typically less than 0.05, implying that the MRCD strikes a balance between being similar to the MCD for tall data and achieving a well-conditioned estimate in the case of fat data.

5 Real data examples

We illustrate the MRCD on two datasets with low n/p , so using the original MCD is not indicated.

5.1 Octane data

The octane data set described in Esbensen et al. (1996) consists of near-infrared absorbance spectra with $p = 226$ wavelengths collected on $n = 39$ gasoline samples. It is known that the samples 25, 26, 36, 37, 38 and 39 are outliers which contain added ethanol (Hubert et al., 2005). Of course, in most applications the number of outliers is not known in advance hence it is not obvious to set the subset size h . The choice of h matters because increasing h improves the efficiency at uncontaminated data but hurts the robustness to outliers. Our recommended default choice is $h = \lceil 0.75n \rceil$, safeguarding the MRCD covariance estimate against up to 25% of outliers.

Alternatively, one could employ a data-driven approach to select h . It consists of computing the MRCD for a range of h values, and looking for an important change in the objective function or the estimates at some value of h . This is not too hard, since we only need to obtain the initial estimates \mathbf{S}^i once. Figure 1 plots the MRCD objective function (10) for each value of h , while Figure 2 shows the Frobenius distance between the values of the MRCD scatter estimates of the standardized data (*i.e.*, $\rho \mathbf{I} + (1 - \rho)\mathbf{S}^*(H_{MRCD})$, as defined in (12)) obtained for $h - 1$ and h . Both figures clearly indicate that there is an important change at $h = 34$, so we choose $h = 33$. The total computation time to produce these plots was only 12 seconds on an Intel(R) Core(TM) i7-5600U CPU with 2.60 GHz.

We then calculate the MRCD estimator with $h = 33$, yielding $\rho = 0.1149$. The condition number of the scatter matrix equals 720. Figure 3 shows the corresponding robust distances

$$RD(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \mathbf{m}_{MRCD})' \mathbf{K}_{MRCD}^{-1} (\mathbf{x}_i - \mathbf{m}_{MRCD})} \quad (21)$$

where \mathbf{m}_{MRCD} and \mathbf{K}_{MRCD} are the MRCD location and scatter estimates of (12). The flagged outliers (red triangles) stand out, showing the MRCD has correctly identified the 6 samples with added ethanol.

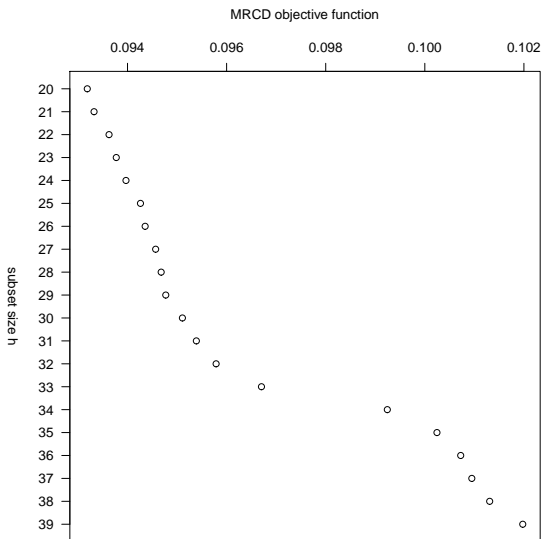


Figure 1: Octane data: MRCD objective value (10) for different values of h .

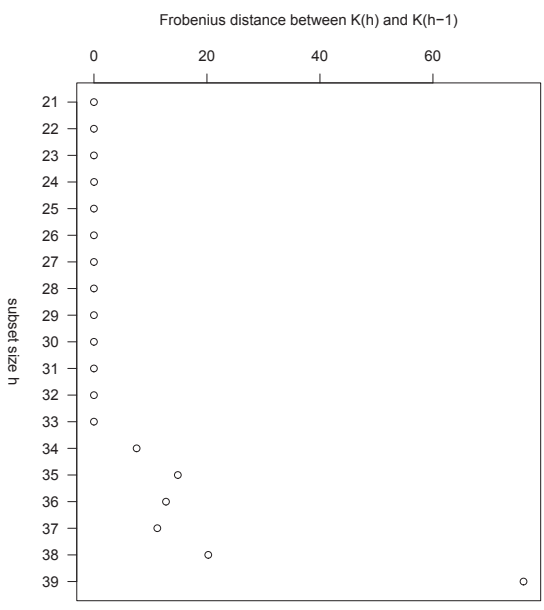


Figure 2: Octane data: Frobenius distance between scatter estimates for $h - 1$ and h .

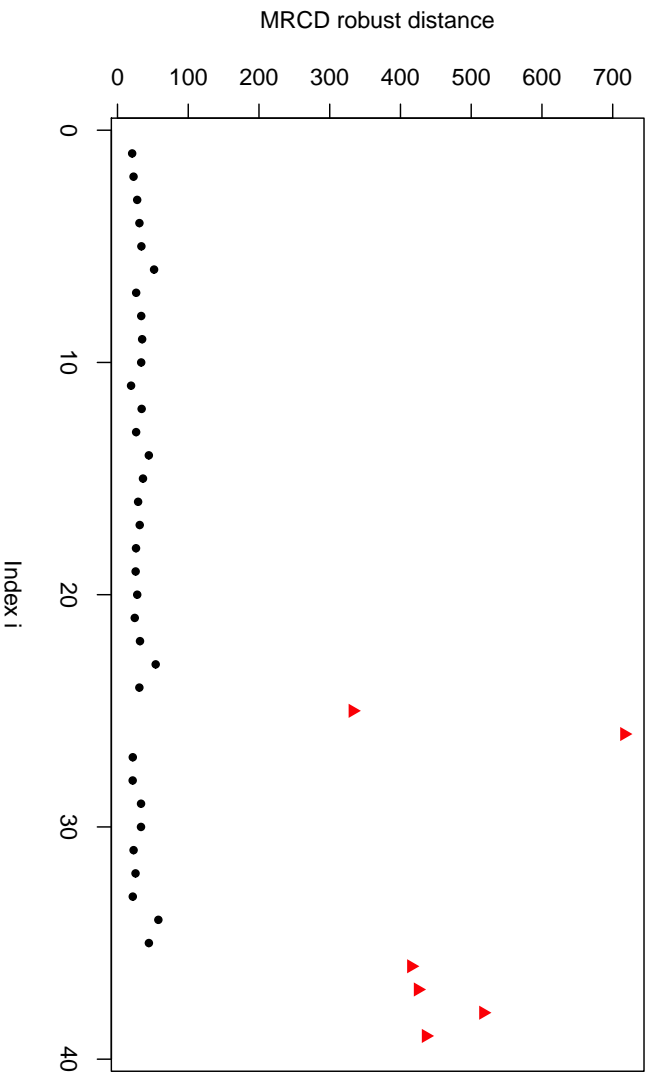


Figure 3: Robust distances of the octane data, based on the MRCD with $h = 33$.

5.2 Murder rate data

Khan et al. (2007) regress the murder rate per 100,000 residents in the $n = 50$ states of the US in 1980 on 25 demographic predictors, and mention that graphical tools reveal one clear outlier. The data can be retrieved from <http://users.ugent.be/~svaelst/software/RLARS.html>.

For lower-dimensional data, Rousseeuw et al. (2004) applied the MCD estimator to the response(s) and predictors together to robustly estimate a multivariate regression. Here we investigate whether for high-dimensional data the same type of analysis can be carried out based on the MRCD. In the murder rate data this yields a total of 26 variables.

As for the octane data, we compute the MRCD estimates for the candidate range of h . In Figure 4 we see a big jump in the objective function when going from $h = 49$ to $h = 50$. But in the plot of the Frobenius distance between successive MRCD scatter matrices (Figure 5) we see evidence of four outliers, which lead to a substantial change in the MRCD when included in the subset.

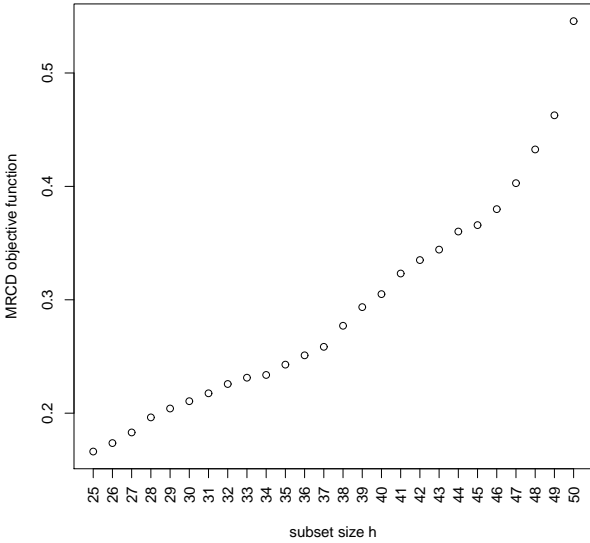


Figure 4: Murder rate data: MRCD objective value (10) for different values of h .

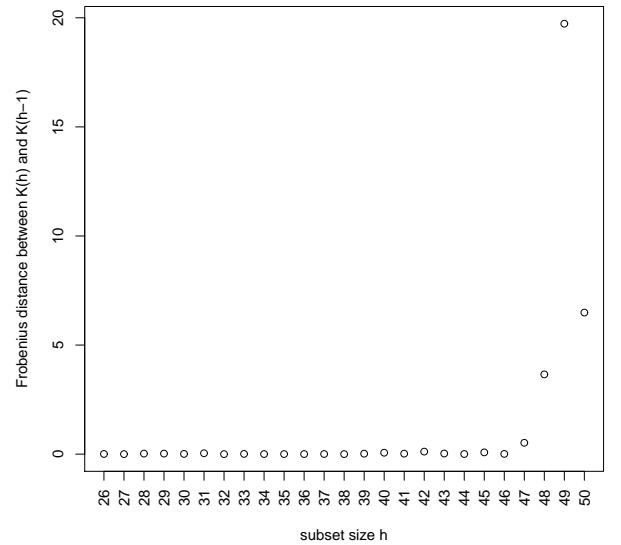


Figure 5: Murder rate data: distance between scatter estimates for $h - 1$ and h .

As a conservative choice we set $h = \lceil 0.9n \rceil = 45$, which allows for up to 5 outliers. We then partition the MRCD scatter matrix on all 26 variables as follows:

$$\mathbf{K}_{MRCD} = \begin{pmatrix} \mathbf{K}_{xx} & \mathbf{K}_{xy} \\ \mathbf{K}_{xy} & \mathbf{K}_{yy} \end{pmatrix},$$

where x stands for the vector of predictors and y is the response variable. The resulting estimate of the slope vector is then

$$\hat{\beta}_{MRCD} = \mathbf{K}_{xx}^{-1} \mathbf{K}_{xy} .$$

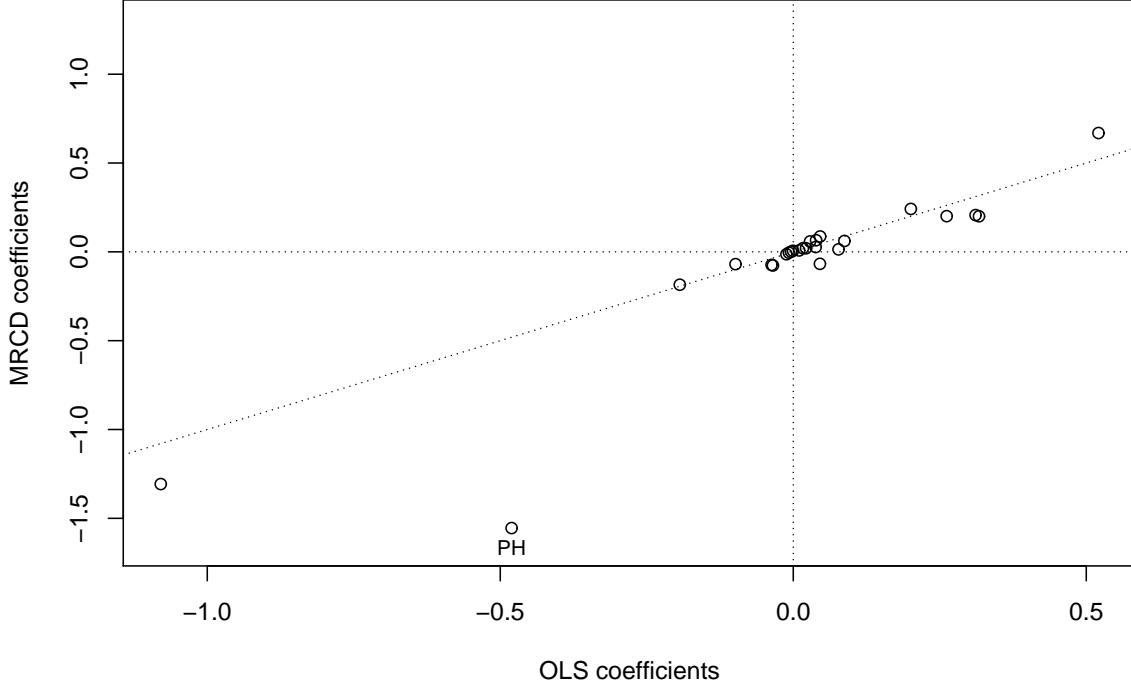


Figure 6: MRCD slopes versus OLS slopes of regressing the murder rate on demographic variables.

The resulting MRCD and OLS slope coefficients are shown in Figure 6. For most variables the coefficients are similar, except for the variable “PH” corresponding to the number of telephones per 100 residents, for which the MRCD coefficient (-1.55) is about three times the OLS coefficient (-0.48). The telephone density might serve as a proxy for the technological level of the state. A negative coefficient then indicates that on average the more technologically advanced the state, the lower the murder rate, other things being equal (which they rarely are).

Figure 7 plots the murder rate against the phone density. In this scatter plot, Arkansas and Nevada appear as outliers. Arkansas is a bad leverage point: it has the lowest phone density (in 1980) but an average murder rate. Nevada is a vertical outlier, as it lies above the downward sloping regression line fitting the bulk of the data, meaning that its murder rate is higher than one would expect on the basis of the phone density alone. The red triangles are the observations not included in the MRCD subset. We see that MRCD regression on all predictors has effectively flagged Arkansas and Nevada. Omitting them, as implicitly done in the MRCD regression, has led to a more negative value of this slope.

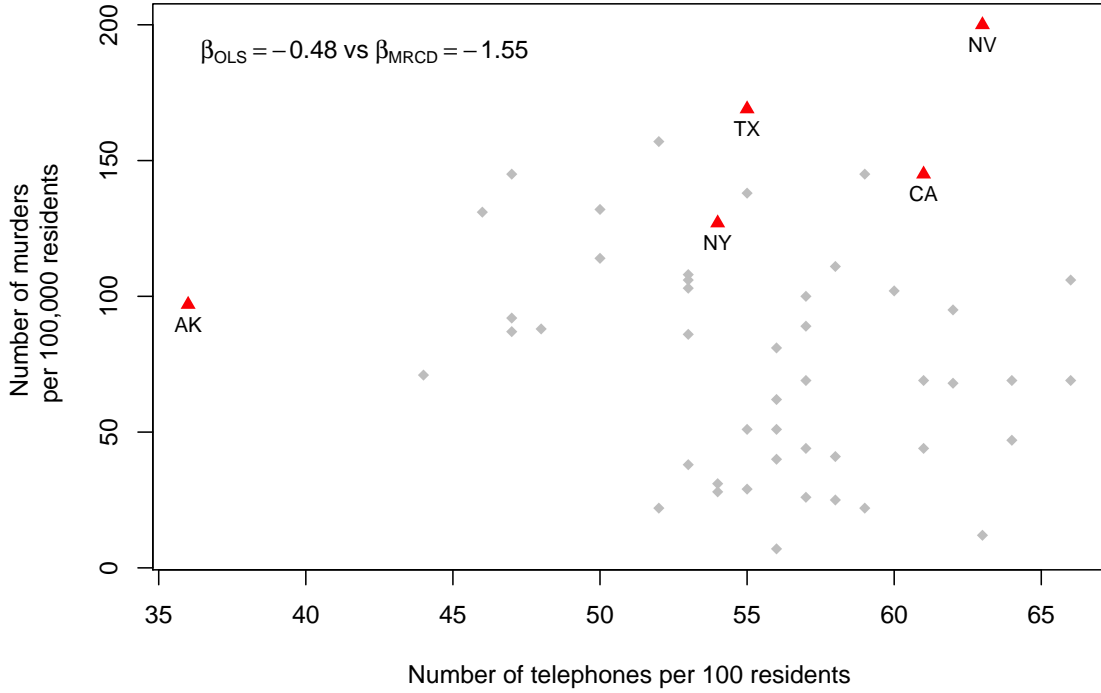


Figure 7: Scatter plot of murder rate per state against phone density. The red triangles indicate the observations that are not included in the final MRCD subset with $h = 45$.

Finally, we note that MRCD regression can be plugged into existing robust algorithms for variable selection, which avoids the limitation mentioned in Khan et al. (2007) that “a robust fit of the *full* model may not be feasible due to the numerical complexity of robust estimation when [the dimension] d is large (e.g., $d \geq 200$) or simply because d exceeds the number of cases, n .” The MRCD could be used in such situations because its computation remains feasible in higher dimensions.

6 Concluding remarks

In this paper we generalize the Minimum Covariance Determinant estimation approach of Rousseeuw (1985) to higher dimensions, by regularizing the sample covariance matrices of subsets before minimizing their determinant. The resulting Minimum Regularized Covariance Determinant (MRCD) estimator is well-conditioned by construction, even when $p > n$, and preserves the good robustness of the MCD. We were able to construct a fast algorithm for the MRCD by generalizing the C-step used by the MCD, and proving that this generalized C-step is guaranteed to reduce the covariance determinant. We verified the performance of the MRCD estimator in an extensive simulation study including both clean and contaminated data. The simulation study also confirms that the MRCD can be interpreted

as a generalization of the MCD, because when n is sufficiently large compared to p and the MCD is well-conditioned the regularization parameter in MRCD becomes zero so the MRCD estimate coincides with the MCD. Finally, we illustrated the use of the MRCD for outlier detection and robust regression on two fat data applications from chemistry and criminology, for which $p > n/2$.

We believe that the MRCD is a valuable addition to the tool set for robust multivariate analysis, especially in high dimensions. We look forward to further research on its use in principal component analysis where the original MCD has proved useful (Croux and Haesbroeck, 2000; Hubert et al., 2005), and analogously in factor analysis (Pison et al., 2003), classification (Hubert and Van Driessen, 2004), clustering (Hardin and Roche, 2004), multivariate regression (Rousseeuw et al., 2004), penalized maximum likelihood estimation (Croux et al., 2012) and other multivariate techniques. A further research topic is to study the finite sample distribution of the robust distances computed from the MRCD. Our experiments have shown that the usual chi-square and F-distribution results for the MCD distances (Hardin and Roche, 2005) are no longer good approximations when p is large relatively to n . A better approximation would be useful for improving the accuracy of the MRCD by reweighting.

References

- Agostinelli, C., A. Leung, V. Yohai, and R. Zamar (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test* 24(3), 441–461.
- Agulló, J., C. Croux, and S. Van Aelst (2008). The multivariate least trimmed squares estimator. *Journal of Multivariate Analysis* 99, 311–338.
- Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics* 22(1), 107–111.
- Boudt, K., J. Cornelissen, and C. Croux (2012). Jump robust daily covariance estimation by disentangling variance and correlation components. *Computational Statistics & Data Analysis* 56(11), 2993–3005.
- Butler, R., P. Davies, and M. Jhun (1993). Asymptotics for the Minimum Covariance Determinant estimator. *The Annals of Statistics* 21(3), 1385–1400.
- Cator, E. and H. Lopuhaä (2012). Central limit theorem and influence function for the MCD estimator at general multivariate distributions. *Bernoulli* 18(2), 520–551.

- Croux, C. and C. Dehon (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications* 19(4), 497–515.
- Croux, C., S. Gelper, and G. Haesbroeck (2012). Regularized Minimum Covariance Determinant estimator. *Mimeo*.
- Croux, C. and G. Haesbroeck (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis* 71(2), 161–190.
- Croux, C. and G. Haesbroeck (2000). Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* 87, 603–618.
- Esbensen, K., T. Midtgaard, and S. Schönkopf (1996). *Multivariate Analysis in Practice: A Training Package*. Camo As.
- Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* 147, 186–197.
- Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(2), 432–441.
- Gnanadesikan, R. and J. Kettenring (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* 28, 81–124.
- Grübel, R. (1988). A minimal characterization of the covariance matrix. *Metrika* 35(1), 49–52.
- Hardin, J. and D. Rocke (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis* 44, 625–638.
- Hardin, J. and D. Rocke (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics* 14(4), 928–946.
- Hubert, M., P. Rousseeuw, and S. Van Aelst (2008). High breakdown robust multivariate methods. *Statistical Science* 23, 92–119.
- Hubert, M., P. Rousseeuw, and K. Vanden Branden (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics* 47, 64–79.
- Hubert, M., P. Rousseeuw, and T. Verdonck (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics* 21(3), 618–637.

- Hubert, M. and K. Van Driessen (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis* 45, 301–320.
- Khan, J., S. Van Aelst, and R. H. Zamar (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association* 102(480), 1289–1299.
- Lopuhaä, H. and P. Rousseeuw (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics* 19, 229–248.
- Maronna, R. and R. H. Zamar (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* 44(4), 307–317.
- Öllerer, V. and C. Croux (2015). Robust high-dimensional precision matrix estimation. In *Modern Nonparametric, Robust and Multivariate Methods*, pp. 325–350. Springer.
- Pison, G., P. Rousseeuw, P. Filzmoser, and C. Croux (2003). Robust factor analysis. *Journal of Multivariate Analysis* 84, 145–172.
- Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association* 79(388), 871–880.
- Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (Eds.), *Mathematical Statistics and Applications, Vol. B*, pp. 283–297. Reidel Publishing Company, Dordrecht.
- Rousseeuw, P. and C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88(424), 1273–1283.
- Rousseeuw, P., C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler (2012). Robustbase: Basic Robust Statistics. R package version 0.92-3.
- Rousseeuw, P., S. Van Aelst, K. Van Driessen, and J. Agulló (2004). Robust multivariate regression. *Technometrics* 46, 293–305.
- Rousseeuw, P. and K. Van Driessen (1999). A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics* 41, 212–223.
- Rousseeuw, P. and B. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85(411), 633–639.

SenGupta, A. (1987). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *Journal of Multivariate Analysis* 23(2), 209–219.

Sherman, J. and W. J. Morrison (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics* 21(1), 124–127.

Won, J.-H., J. Lim, S.-J. Kim, and B. Rajaratnam (2013). Condition-number-regularized covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(3), 427–450.

Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum report* 42, 106.

Appendix A: Proof of Theorem 1

Generate a p -variate sample \mathbf{Z} with $p+1$ points for which $\mathbf{\Lambda} = \frac{1}{p+1} \sum_{j=1}^{p+1} (\mathbf{z}_j - \bar{\mathbf{z}})(\mathbf{z}_j - \bar{\mathbf{z}})'$ is nonsingular and $\bar{\mathbf{z}} = \frac{1}{p+1} \sum_{j=1}^{p+1} \mathbf{z}_j$. Then $\tilde{\mathbf{z}}_i = \mathbf{\Lambda}^{-1/2}(\mathbf{z}_i - \bar{\mathbf{z}})$ has mean zero and covariance matrix \mathbf{I}_p . Now compute $\mathbf{y}_i = \mathbf{T}^{1/2} \tilde{\mathbf{z}}_i$, hence \mathbf{Y} has mean zero and covariance matrix \mathbf{T} .

Next, create the artificial dataset

$$\tilde{\mathbf{X}}^1 = (w_1(\mathbf{x}_1^1 - \mathbf{m}_1), \dots, w_h(\mathbf{x}_h^1 - \mathbf{m}_1), w_{h+1}\mathbf{y}_1, \dots, w_k\mathbf{y}_{p+1})$$

with $k = h + p + 1$ points, where $\mathbf{x}_1^1, \dots, \mathbf{x}_h^1$ are the members of H_1 . The factors w_i are given by

$$w_i = \begin{cases} \sqrt{k(1-\rho)/h} & \text{for } i = 1, \dots, h \\ \sqrt{k\rho/(p+1)} & \text{for } i = h+1, \dots, k \end{cases}.$$

The mean and covariance matrix of $\tilde{\mathbf{X}}^1$ are then

$$\frac{1}{k} \sum_{i=1}^k \tilde{\mathbf{x}}_i^1 = \sqrt{\frac{1-\rho}{kh}} \sum_{i=1}^h (\mathbf{x}_i^1 - \mathbf{m}_1) + \sqrt{\frac{\rho}{k(p+1)}} \sum_{j=1}^{p+1} \mathbf{y}_j = 0$$

and

$$\begin{aligned} \frac{1}{k} \sum_{i=1}^k \tilde{\mathbf{x}}_i^1 (\tilde{\mathbf{x}}_i^1)' &= \frac{1-\rho}{h} \sum_{i=1}^h (\mathbf{x}_i^1 - \mathbf{m}_1)(\mathbf{x}_i^1 - \mathbf{m}_1)' + \frac{\rho}{p+1} \sum_{j=1}^{p+1} \mathbf{y}_j \mathbf{y}_j' \\ &= (1-\rho)\mathbf{S}_1 + \rho\mathbf{T} = \mathbf{K}_1. \end{aligned}$$

The regularized covariance matrix \mathbf{K}_1 is thus the actual covariance matrix of the combined data set $\tilde{\mathbf{X}}^1$. Analogously we construct

$$\tilde{\mathbf{X}}^2 = (w_1(\mathbf{x}_1^2 - \mathbf{m}_2), \dots, w_h(\mathbf{x}_h^2 - \mathbf{m}_2), w_{h+1}\mathbf{y}_1, \dots, w_k\mathbf{y}_{p+1})$$

where $\mathbf{x}_1^2, \dots, \mathbf{x}_h^2$ are the members of H_2 . $\tilde{\mathbf{X}}_2$ has zero mean and covariance matrix $\mathbf{K}_2 = (1-\rho)\mathbf{S}_2 + \rho\mathbf{T}$.

Denote $d_{\mathbf{K}_1}(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}'(\mathbf{K}_1)^{-1}\tilde{\mathbf{x}}$. We can then prove that:

$$\frac{1}{k} \sum_{i=1}^h d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^2) = \frac{1-\rho}{h} \sum_{i=1}^h d_{\mathbf{K}_1}(\mathbf{x}_i^2 - \mathbf{m}_2) \quad (22)$$

$$\leq \frac{1-\rho}{h} \sum_{i=1}^h d_{\mathbf{K}_1}(\mathbf{x}_i^2 - \mathbf{m}_1) \quad (23)$$

$$\leq \frac{1-\rho}{h} \sum_{i=1}^h d_{\mathbf{K}_1}(\mathbf{x}_i^1 - \mathbf{m}_1) \quad (24)$$

$$= \frac{1}{k} \sum_{i=1}^h d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^1) \quad (25)$$

in which the second inequality (24) is the condition (19).

The first inequality (23) can be shown as follows. Put $\mathbf{z}_i = (\mathbf{K}_1)^{-1/2}\mathbf{x}_i^2$ and $\tilde{\mathbf{z}} = (\mathbf{K}_1)^{-1/2}\mathbf{m}_1$ and note that $\bar{\mathbf{z}} = (\mathbf{K}_1)^{-1/2}\mathbf{m}_2$ is the average of the \mathbf{z}_i . Then (23) becomes

$$\sum_{i=1}^h \|\mathbf{z}_i - \bar{\mathbf{z}}\|^2 \leq \sum_{i=1}^h \|\mathbf{z}_i - \tilde{\mathbf{z}}\|^2,$$

which follows from the fact that $\tilde{\mathbf{z}}$ is the unique minimizer of the least squares objective $\sum_{i=1}^k \|\mathbf{z}_i - c\|^2$, so (23) becomes an equality if and only if $\tilde{\mathbf{z}} = \bar{\mathbf{z}}$ which is equivalent to $\mathbf{m}_2 = \mathbf{m}_1$.

It follows that

$$\begin{aligned} \sum_{i=1}^k d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^2) &= \sum_{i=1}^h d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^2) + \frac{\rho}{p+1} \sum_{j=1}^{p+1} d_{\mathbf{K}_1}(\mathbf{y}_j) \\ &\leq \sum_{i=1}^h d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^1) + \frac{\rho}{p+1} \sum_{j=1}^{p+1} d_{\mathbf{K}_1}(\mathbf{y}_j) \\ &= \sum_{i=1}^k d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^1) . \end{aligned}$$

Now put

$$b = \frac{\sum_{i=1}^k d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^2)}{\sum_{i=1}^k d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^1)} \leq 1 .$$

If we now compute distances relative to $b\mathbf{K}_1$, we find

$$\begin{aligned}
\frac{1}{k} \sum_{i=1}^k d_{b\mathbf{K}_1}(\tilde{\mathbf{x}}_i^2) &= \frac{1}{b} \frac{1}{k} \sum_{i=1}^k d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^2) = \frac{1}{k} \sum_{i=1}^k d_{\mathbf{K}_1}(\tilde{\mathbf{x}}_i^1) = \frac{1}{k} \sum_{i=1}^k (\tilde{\mathbf{x}}_i^1)' (\mathbf{K}_1)^{-1} \tilde{\mathbf{x}}_i^1 \\
&= \frac{1}{k} \sum_{i=1}^k (\mathbf{K}_1^{-1/2} \tilde{\mathbf{x}}_i^1)' (\mathbf{K}_1^{-1/2} \tilde{\mathbf{x}}_i^1) = \text{Trace} \left(\frac{1}{k} \sum_{i=1}^k (\mathbf{K}_1^{-1/2} \tilde{\mathbf{x}}_i^1)' (\mathbf{K}_1^{-1/2} \tilde{\mathbf{x}}_i^1) \right) \\
&= \text{Trace} \left((\mathbf{K}_1)^{-1/2} \left(\frac{1}{k} \sum_{i=1}^k (\tilde{\mathbf{x}}_i^1) (\tilde{\mathbf{x}}_i^1)' \right) (\mathbf{K}_1)^{-1/2} \right) = \text{Trace}(\mathbf{I}_p) = p .
\end{aligned}$$

From the theorem in Grübel (1988), it follows that \mathbf{K}_2 is the unique minimizer of $\det(\mathbf{S})$ among all \mathbf{S} for which $\frac{1}{k} \sum_{i=1}^k d_{\mathbf{S}}(\tilde{\mathbf{x}}_i^2) = p$ (note that the mean of $\tilde{\mathbf{x}}_i^2$ is zero). Therefore

$$\det(\mathbf{K}_2) \leq \det(b\mathbf{K}_1) \leq \det(\mathbf{K}_1) .$$

We can only have $\det(\mathbf{K}_2) = \det(\mathbf{K}_1)$ if both of these inequalities are equalities. For the first, by uniqueness we can only have equality if $\mathbf{K}_2 = b\mathbf{K}_1$. For the second inequality, equality holds if and only if $b = 1$. Combining both yields $\mathbf{K}_2 = \mathbf{K}_1$. Moreover, $b = 1$ implies that (23) becomes an equality, hence $\mathbf{m}_2 = \mathbf{m}_1$. This concludes the proof of Theorem 1.

Appendix B: The OGK estimator

Maronna and Zamar (2002) presented a general method to obtain positive definite and approximately affine equivariant robust scatter matrices starting from a robust bivariate scatter measure. This method was applied to the bivariate covariance estimate of Gnanadesikan and Kettenring (1972). The resulting multivariate location and scatter estimates are called orthogonalized Gnanadesikan-Kettenring (OGK) estimates and are calculated as follows:

1. Let $m(\cdot)$ and $s(\cdot)$ be robust univariate estimators of location and scale.
2. Construct $\mathbf{y}_i = \mathbf{D}^{-1} \mathbf{x}_i$ for $i = 1, \dots, n$ with $\mathbf{D} = \text{diag}(s(X_1), \dots, s(X_p))$.
3. Compute the ‘pairwise correlation matrix’ \mathbf{U} of the variables of $\mathbf{Y} = (Y_1, \dots, Y_p)$, given by $u_{jk} = 1/4(s(Y_j + Y_k)^2 - s(Y_j - Y_k)^2)$. This \mathbf{U} is symmetric but not necessarily positive definite.
4. Compute the matrix \mathbf{E} of eigenvectors of \mathbf{U} and
 - (a) project the data on these eigenvectors, i.e. $\mathbf{V} = \mathbf{Y}\mathbf{E}$;

(b) compute ‘robust variances’ of $\mathbf{V} = (V_1, \dots, V_p)$, i.e. $\mathbf{\Lambda} = \text{diag}(s^2(V_1), \dots, s^2(V_p))$;

(c) set the $p \times 1$ vector $\hat{\boldsymbol{\mu}}(\mathbf{Y}) = \mathbf{E}\mathbf{m}$ where $\mathbf{m} = (m(V_1), \dots, m(V_p))^T$, and compute the positive definite matrix $\hat{\boldsymbol{\Sigma}}(\mathbf{Y}) = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^T$.

5. Transform back to \mathbf{X} , i.e. $\hat{\boldsymbol{\mu}}_{\text{OGK}} = \mathbf{D}\hat{\boldsymbol{\mu}}(\mathbf{Y})$ and $\hat{\boldsymbol{\Sigma}}_{\text{OGK}} = \mathbf{D}\hat{\boldsymbol{\Sigma}}(\mathbf{Y})\mathbf{D}^T$.

In the OGK algorithm $m(\cdot)$ is a weighted mean and $s(\cdot)$ is a 1-step M-estimator of scale. Step 2 makes the estimate location invariant and scale equivariant, whereas the next steps replace the eigenvalues of \mathbf{U} (some of which may be negative) by positive numbers. (Should some of the $s^2(V_i)$ be zero, we could replace them by small positive numbers.)